International Journal of Industrial Ergonomics 39 (2009) 621–627

Contents lists available at ScienceDirect

International Journal of Industrial Ergonomics

journal homepage: www.elsevier.com/locate/ergon

Web evaluation: Heuristic evaluation vs. user testing

Wei-siong Tan [a], Dahai Liu [b], Ram Bishu [a,*]

[a] Department of Industrial and Management Systems Engineering, University of Nebraska at Lincoln, Room 175, NH, Lincoln 68588, USA
[b] Department of Human Factors and Systems, Embry-Riddle Aeronautical University, 600 S. Clyde Morris Blvd., Daytona Beach, FL 32114, USA

ARTICLE INFO

Article history:
Received 31 May 2007
Received in revised form
25 January 2008
Accepted 14 February 2008
Available online 23 May 2008

Keywords:
Usability testing
Heuristic evaluation

ABSTRACT

It is very important that designers recognize the benefits and limitations of different usability inspection methods. This is because the quality of the usability evaluation is dependent on the method used. Two of the most popular usability evaluation techniques are user testing and heuristic analysis. The main objective of this study was to compare the efficiency and effectiveness between user testing and heuristic analysis in evaluating four different commercial web sites. The results showed that both user testing and heuristic analysis addressed different usability problems. Analysis by severity of problems found and diminishing return analysis model on the relationship between the number of new problems discovered with users and evaluators used showed that both methods are equally efficient and effective in addressing different categories of usability problems. These significant differences found between these two methods suggested that the two methods are complimentary and should not be competing. In order for better evaluation results, both user testing and heuristic analysis are still needed.

Relevance to industry: The research findings from this study will be of particular value to the web development industry and communities. Knowledge regarding the differences between user testing and heuristic evaluation will enable appropriate business decisions to be made on when and how to apply these methods to improve the overall efficiency of the design process.

Published by Elsevier B.V.

1. Introduction

On the web, usability plays a crucial role as user experience is emphasized above anything else and it is an inherent design characteristic and requirement, and closely related to various user errors (Rexfelt and Rosenblad, 2006; Latorella and Prabhu, 2000; Gramopadhye and Drury, 2000). Thus, it is essential to create a well-designed web site that is highly usable. The question is one, of deciding what constitutes a well-designed site and how to evaluate the same? Different usability evaluation techniques have been developed and incorporated into the design and development of web sites. Among these techniques, user testing and heuristic analysis are perhaps two of the most popular ones.

1.1. Comparison of two methods

Both user testing and heuristic evaluation methods provide valuable insight of usability problems in both 'finished-ready-to-launch-interface' as well as in 'iterative design-construction phase of an interface'. User testing relies mainly on the experience and comments of the users and is usually conducted in a scenario-based environment. As a result, user testing would usually evaluate according to what already exists, rather than to what is possible. On the other hand, heuristic analysis relies mainly on the expertise and knowledge of human factors engineers that would evaluate the web site based on a set of heuristics. Both of these methods have their individual strengths and weaknesses, and neither one guarantees an optimal result. The issue of whether these methods are complementary or competing, has been the topic of considerable research.

A number of published studies have compared these evaluation methods. Liljegren and Osvalder (2004) investigated efficiencies of several types of usability evaluation tools. They found that a particular tool is most efficient for a specific aspect of the user interface and for a certain sequence, for example, user testing works better when a cognitive walkthrough is done and certain trials have been performed. Jeffries et al. (1991) had found that heuristic analysis discovered approximately three times more problems than user testing. Heuristic analysis found 105 problems while user testing found 31 problems. However, Jeffries et al. (1991) reported that more severe problems were discovered through user testing, as compared to heuristic analysis. Liljegren (2006) investigated four common methods including hierarchical task analysis, cognitive walkthrough, heuristic evaluation and user testing, based on their thoroughness, validity, reliability, cost effectiveness and

* Corresponding author. Tel.: +1 402 4722393; fax: +1 402 472 1384.
E-mail address: rbishu@unlnotes.unl.edu (R. Bishu).

0169-8141/$ – see front matter Published by Elsevier B.V.
doi:10.1016/j.ergon.2008.02.012

clarity. It has been found that user testing is recommended to be the primary method in usability evaluations, as they fulfill the criteria and address the 'difficulty to make errors' aspect of overall usability. Lindgaard (2006) has also found similar results. Nielsen and Mack (1994) used 11 evaluators for heuristic analysis and 4 subjects for user testing to find usability problems in a complex telephone company application. The two methods found 17 common problems while heuristic analysis found 23 problems unique to itself and user testing found 4 problems unique to it. Doubleday et al. (1997) had used heuristic analysis and user testing to evaluate an information retrieval interface. A total of 5 evaluators and 20 subjects participated in this study. Approximately, 39% of the usability problems were discovered by user testing, while 40% were unique to heuristic analysis. The rest were common to both methods. Desurvire et al. (1992) had used different groups of experts to perform heuristic analysis. User testing was used as a bench mark to assess the total number of problems and evaluated heuristic analysis against this. Desurvire et al. (1992) reported that heuristic analysis predicted 44% of the total problems and 29% of the severe problems. There is an issue with user testing being used as a bench mark to compare the ability of heuristic analysis (Desurvire et al., 1992). The power of "seeing is believing" creates a bias for software developers towards user testing (Nielsen and Mack, 1994), as usability problems that are not detected by user testing are considered to be "false positive" problems. Not all usability problems are detected in user testing, and some of the missed problems can indeed be very real and costly. Nielsen and Mack (1994) argues that these problems may not have been detected because their impact may have been too short a duration to be observed, and some problems may occur too infrequently to be observed by small groups of users that were tested. Users may also demonstrate an ability to perform the task, in spite of the interface error (Doubleday et al., 1997). Further diversity of users of web sites across the globe, prohibits user testing by all segments of global users. As a result, usability problems that were deemed to be "false positive" problems may be real problems in actual usage.

Studies also reported differences in the type of usability problems that were discovered by both the usability inspection methods. Jeffries et al. (1991) defined three different types of usability categories, which include consistency, recurring, and general problems. However, these three categories, alone, fail to represent all the common types of problems encountered by a typical user interface. It is also relevant to mention here that problems differ in severity. Some of the problems may be superficial and frustrating at best, while others may be functionally debilitating.

1.2. Number and type of evaluators

Number of evaluators used in different studies varied. For example, Nielsen and Mack (1994) used three times more subjects for heuristic evaluation as compared to user testing. On the other hand, Doubleday et al. (1997) used more people for user testing than for heuristic evaluation. Although, Jeffries et al. (1991) used roughly an equal number of users and evaluators, the sample size was relatively small, and the conclusions drawn were interface specific and lacked generalizability.

Nielsen and Landauer (1993) were the first to develop a relationship between the number of subjects/evaluators and the number of problems found. They showed that the number of usability problems found in both user testing and heuristic analysis with n users is

$$\text{ProblemsFound}\ (i) = N(1-(1-L)^n)$$

where ProblemsFound (i) is the number of different usability problems found by aggregating reports from i independent evaluators, N is the total number of usability problems in the design, L is the proportion of usability problems discovered while testing with a single user, and n is the number of users.

From the study, Nielsen and Landauer (1993) claim that 5 users are enough to catch 85% of the problems on practically all web sites. Nielsen and Landauer (1993) also reported that as the number of evaluators increased, the additional problems discovered per evaluator decreased. The study has been accepted as an industry standard (Nielsen and Landauer, 1993). However, Spool and Schroeder (2001) have reported different results. They conducted an user testing on a web site with 49 users, who were asked to perform a single task, i.e., to purchase a product. The result of the study revealed that five users could only find 35% of the problems. Woolrych and Cockton (2002) also reached the conclusion that 5 users are not adequate.

For heuristic analysis, Nielsen and Mack (1994) recommends using 3–5 evaluators if "single expert" usability specialists were utilized. They recommend using 2–3 evaluators if "double expert" usability specialists were used. Double experts found 60% of the usability problems, making them 2.7 times as good as novices and 1.5 times as good as single experts. A "double expert" evaluator is defined as a person with a usability background and a specific application area. A "single expert" evaluator is defined as a person with general usability experiences. The level of expertise of the evaluators is a very important factor, as there is a systematic group difference in evaluator performance in addition to individual performance (Nielsen and Mack, 1994; Nielsen and Molich, 1990). Thus better results would be obtained if the right group of evaluators were utilized. Also there is considerable amount of individual differences in problems, identified by experts, in heuristic evaluation. In fact the comparative usability evaluation (CUE) method advocated by Molich et al. (1998) uses these differences to improve end use customer applications.

In summary, considerable confusion exists in literature on the relative effectiveness of user testing and heuristic evaluation of web evaluation. It is a fact that these methods identify more unique problems than common ones. The number of problems identified by both these methods should be greater than either alone. It is possible that they may be complementary instead of competing. It is clear, however, that the number of evaluators, number of subjects, types of evaluators and subjects, and scenarios used, all influence the evaluation process. Further, development of a web site involves number of stages, starting from concept design to final construction and launch. Published studies have not looked at relative usefulness of user testing and heuristic analysis at different design and development stages. The evaluation needs are different at these stages. Earlier stages in the design may warrant quick and easy methods that give immediate feedback, while later stages with more details added to the design and high fidelity prototype available may warrant scenario-based user testing (Preece et al., 2002).

The primary purpose of the present study was to compare user testing and heuristic analysis. The intent was also to compare the quantity, severity, and type of usability problems discovered by both methods. The final objective was to develop functional relationship between the number of new usability problems found and the number of users or evaluators used.

2. Method

2.1. The interface

A total of four existing commercial web sites were evaluated for this study, the first two web sites were considered to have an average

number of usability problems, while the other two were considered to have a high number of usability problems. The rationale for this was to have a reasonable number of usability problems that could facilitate comparison of the two evaluation methods. Typically web sites, in their earlier stages of development, are expected to have larger number of usability problems than those at later stages of development. A screening procedure was used to select the web sites, from a candidate list of potential sites. The procedure involved scoring a test list of sites on practical usability criteria by five practitioners. The top two and the bottom two were then selected (Tan, 2003). These are referred as sites A1, A2, B1, and B2 ,respectively, hereafter.

The first web site provided information, reviews, and ratings to its members on various consumer products. The second web site primarily sold nutrition supplements and vitamins. The third web site sold computer, software, and networking products. It also provided financing programs for the products that they sold. The fourth web site sold rare and collectable watches.

2.2. Subjects

A total of 12 users were recruited for user testing and 9 evaluators were recruited for heuristic analysis for each of the web sites. Users for user testing were recruited based on a profile that was established by surveying a representative sample of the user population. These users were non-experts and non-power users, which means they have not had any web evaluation experiences but with some experience in surfing the web.

Experts were recruited for performing heuristic analysis. For this study an expert was defined as one, who had graduate level coursework in human computer interaction, and in human factors of web design, who had already been educated and participated in at least one heuristic web evaluation project. This is consistent with the notion that expert evaluators should be used for heuristic evaluation, as they provide better results (Desurvire et al., 1992).

2.3. Scenarios

Typically user testing is driven by scenarios-based tasks that users need to perform, while heuristic analysis is driven by the exploration and evaluation of the web sites by evaluators as they see fit. Heuristic analysis evaluators will decide on their own as to how they want to proceed with evaluating the interface (Nielsen and Mack, 1994). A total of 5 scenarios that represented typical site usage situations in real life were given to both the users and evaluators. Scenarios are similar across the web sites, the following list illustrates a sample scenario list for one web site:

- Scenario 1—registration on the site and creating a notebook.
- Scenario 2—researching information on a specific automobile.
- Scenario 3—use of tools.
- Scenario 4—researching information on a consumer product.
- Scenario 5—exit the site–access saved payment information.

These scenarios for user testing detailed the task that users were required to perform, while the scenarios for heuristic analysis were open ended and went to the extent of highlighting the areas that evaluators would need to evaluate. The purpose of keeping the scenarios same for both methods of evaluation was to facilitate comparison across these methods.

2.4. Procedure

2.4.1. User testing procedure

The experiment included three sessions: the planning session, the testing session, and the reporting session (Tan, 2003). During the planning stage, the subjects explained the testing procedure using a set of training scenarios. They were also explained about the post-evaluation questionnaires, which they were supposed to respond to, after the test. The post-evaluation questionnaire dealt with users' general impression of the web site, usage of terminology, information content, and information structure. During the testing, problems and feedbacks from users were recorded. In the post-testing session, users were given the opportunity to provide feedback and opinion regarding problems that they had faced during the test. It also served as an opportunity for the observer to clarify any doubts that they might have had during the test with regard to the observations made. In the reporting stage, the inherent problems and inconsistencies according to post-evaluation questionnaires, interviews, and expert discussions were identified.

2.4.2. Heuristics analysis procedure

The evaluators independently examined the interfaces and judged their compliance with a set of heuristics. Most of the heuristics suggested by Nielsen and Mack (1994) was used and all evaluators are familiar with these heuristics. Each evaluator worked through the interface at least twice. The first time to get a feel for the flow of the interaction, and the second time to focus on the specific interface elements within the context of the larger whole (Nielsen and Mack, 1994). They were also required to be as specific as possible and to list each usability problem separately. After the completion of heuristic analysis a debriefing session was held where the findings of the study were reported and compiled.

2.5. Design

A mixed factorial design was used. The independent variables were: (a) method of evaluation, (b) type of site, and (c) scenarios. The dependent variables were performance time, number, type, and severity of problems.

3. Results

Data were analysed using analyses of variance (ANOVA). Table 1 shows the ANOVA summary on number of problems found.

While the web sites were expected to show different amount of usability problems (they were selected on that basis), the significance of evaluation method was interesting. Among the usability problems found by the evaluation methods being tested here, a small proportion was found by both, while a large proportion of problems were different and method specific. Table 2 gives the total number of problems identified by the two methods. Common problems refer to intersection of problems found by the two methods, and percentage refers to the percentage of problems divided by the total number of problems. From these, it can be noted that:

- Heuristic analysis discovered about 60% of the problems, while user testing discovered 30% of the problems. The remaining 10% of the problems were discovered by both the methods.
- The above findings were consistent across both the site type.

Table 1
ANOVA summary.

Source	Significance
Web site type	Significant
Evaluation method	Significant
Scenarios	Significant
Web site types*method	Not significant

Table 2
Number of usability problems found by user testing and heuristic analysis.

	Website A1	Website A2	Subtotal for A1 and A2 (percentage)	Web site B1	Web site B2	Subtotal for B1 and B2 (percentage)
Common	5	8	13 (7%)	11	15	26 (11%)
Heuristics	55	46	101 (58%)	78	72	150 (61%)
User	23	38	61 (35%)	45	24	69 (28%)
Total	73	76	149	102	81	183

- Heuristic analysis appeared to identify more number of problems than user testing.

3.1. Analysis by severity of problems

The problems were classified on the basis of severity. Not all usability problems are similar. One would expect problems to have a range of effects on the site usability with some crippling the functionality, while others just being cosmetic. A set of severity criteria was established to rate the severity of the problems. The three different severity ratings included severe, medium, and mild problems.

- Severe problems—includes catastrophic usability problems where users are unable to do their work and major problems where users have difficulty, but are able to find workarounds. Fixing them is mandatory.
- Medium problems—includes medium usability problems where users stumble over the problem, but can quickly adapt to it. Fixing them should be given medium priority.
- Mild problems—includes minor usability problems where users can easily work around the problem. Fixing them should be given low priority.

Fig. 1 shows the severity levels of the problems identified by the two evaluation methods. It appears that the respective proportion of problems of type 'severe', 'medium', and 'mild' is same for the two evaluation methods. Table 3 shows the distribution (in proportion) of problems identified by severity for the good and bad sites. It appears that the respective proportion of problems of type 'severe', 'medium', and 'mild' are similar for the two types of sites as well.

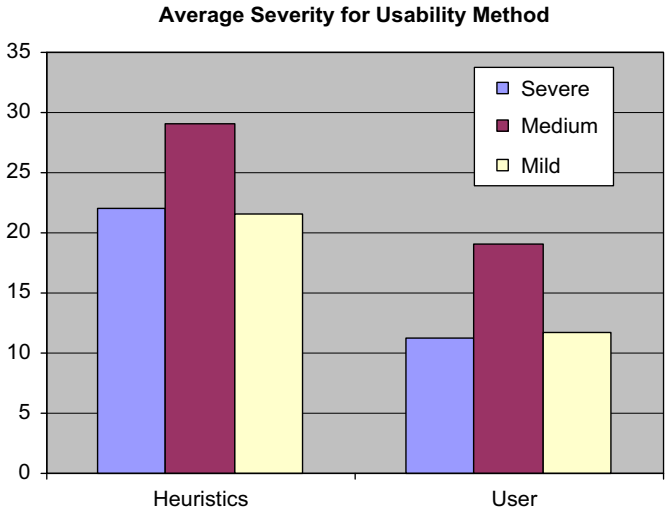

Fig. 1. Distribution of problems by severity by the two evaluation methods (*y*-axis: severity score).

Table 3
Distribution of problems (number of problems found) by severity type identified by the two evaluation methods for good and bad sites.

	Severe	Medium	Mild
Sites A1 and A2—HA	37	28	35
Sites A1 and A2—UT	27	42	31
Sites B1 and B2—HA	34	33	33
Sites B1 and B2—UT	35	32	33

UT=user testing.
HA=heuristic analysis.

3.2. Analysis by site attributes

In designing a site, there are different attributes such as content, navigation, and compatibility that need to be addressed. A total of seven site attributes were developed (Tan, 2003). The next step was to study the distribution of problems with respect to these site attributes. The problems identified were classified into these seven attributes.

- Attribute I—navigation
- Attribute II—compatibility
- Attribute III—information content
- Attribute IV—layout organization and structure
- Attribute V—usability and availability of tools
- Attribute VI—common look and feel
- Attribute VII—security and privacy

A mean comparison analysis was conducted on the number of problems found for the two methods on these seven categories. The analysis reveals that both user testing and heuristic analysis are equally effective in addressing the different usability problems with the exception of attribute II, which address the compatibility issues, and attribute VII, which address the security and privacy issues. User testing had totally failed to address both these issues. All seven usability categories consisted of severe, medium, and mild problems. Attributes II and VII are deemed as more important they consisted of only severe and medium problems. Table 4 illustrates the distribution of problems found for the seven attributes across three severity levels and three types of problems. Fig. 2 shows the distribution of problems identified by the seven attributes. It can be seen that most severe problems are found in attributes 1 and 5, most medium problems found in 1, 3, 5, and 7; and most mild problems found in 1, 3, 4, 5, and 7. So for both evaluation methods, Attribute I (navigation) appears to be the most, followed by attributes III (content), IV (layout), V (usability), and VI (look and feel).

Table 4
Distribution of problems (number of problems) by attributes, severity level and evaluation types.

Severity	Type	Seven attributes						
		I	II	III	IV	V	VI	VII
Severe	Common	11	0	2	4	10	1	5
	Heuristic	37	3	7	8	33	3	15
	User	17	0	11	4	18	1	4
Medium	Common	12	0	32	5	5	5	0
	Heuristic	62	4	14	22	42	17	5
	User	35	0	0	18	14	19	1
Mild	Common	1	0	0	1	0	1	0
	Heuristic	22	0	28	17	10	19	0
	User	21	0	19	11	9	10	0

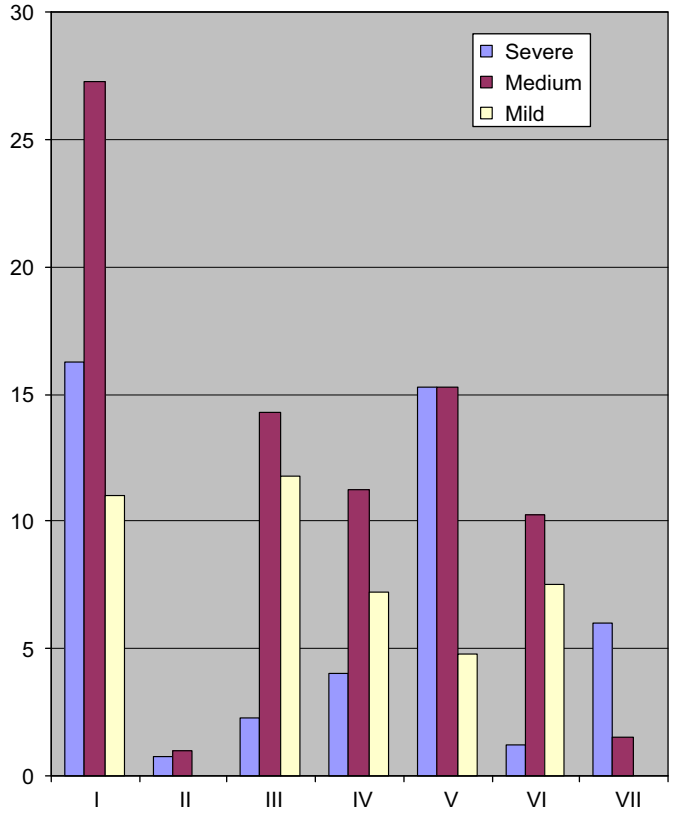

Fig. 2. Distribution of problems identified by Severity and by attributes (*y*-axis: Number of problems found).

Fig. 3 shows a similar plot for the problems identified by the two methods of evaluation across the seven attributes. It is consistent with Figs. 1 and 2, in the sense that (a) heuristic evaluation identified more number of problems, and (b) attributes I, III, IV, V, and VI were the most predominant for both methods of evaluation.

3.3. Cumulative analysis

Fig. 4 shows the plot of cumulative number of problems identified by number of evaluators. The plots do become asymptotic

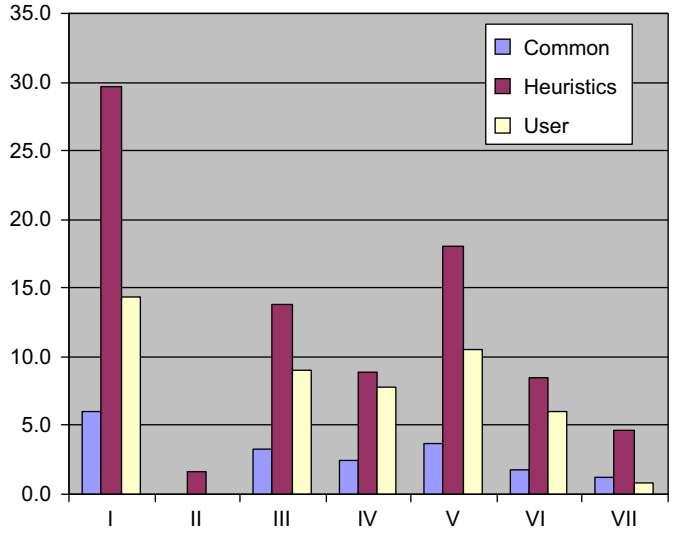

Fig. 3. Distribution of problems identified by method of evaluation and by attributes (*y*-axis: number of problems found).

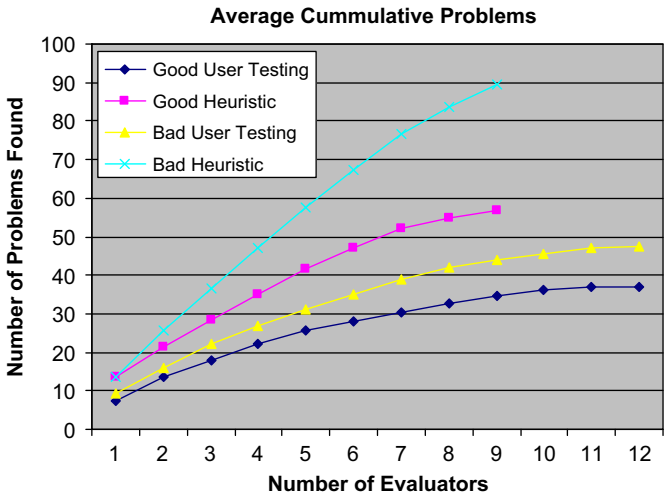

Fig. 4. Plot of cumulative problems and number of evaluators for one bad site and one good site.

after 7 or 8 evaluators. Fig. 4 seems to be very similar in shape to the one reported by Nielsen and Landauer (1993). At this point it was decided to fit an appropriate equation to the data.

3.4. Model development

Previous research has identified that as the number of users or evaluators increase in user testing and/or heuristic analysis, the number of additional new problems found would decrease (Nielsen and Landauer, 1993), that is to say, the marginal number of new problems discovered decreases with increased number of new users and evaluators. This relationship is defined as diminishing return relationship. Both Nielsen and Landauer (1993) and Spool and Schroeder (2001) have shown that there is a diminishing return relationship between the number of new problems discovered with users and evaluators used in a usability study. Table 5 shows the summary of equations, while Fig. 5 shows the plot of the best model.

From Fig. 5, the following observations can be made:

- User testing and heuristic analysis have the same diminishing relationship, if the sample size of the users and evaluators were less than 5. However, it is known at this point that the two methods identify different usability problems.
- Heuristic analysis tends to discover more new problems, for each added evaluator, if more than 5 evaluators are utilized.
- Five evaluators could only find 35% of the usability problems in a user interface. This is different from what Neilson reported in his study (Neilson, 1993). The author reported that 5 evaluators are adequate to identify 75% of the usability problems.

Table 5
Summary of fitted functions.

Method	Equations	R^2 (%)	Preferred
UT	Number of problems=exp(−0.665036−1.80281/Users)	81.98	
UT	Number of problems=1/(1.28751+8.89685/Users)	92.22	*X*
HA	Number of problems=exp(−0.647765−1.86869/Evaluators)	88.84	
HA	Number of problems=1/(0.934244+9.94869/Evaluators)	90.82	*X*

UT=user testing.
HA=heuristic analysis.

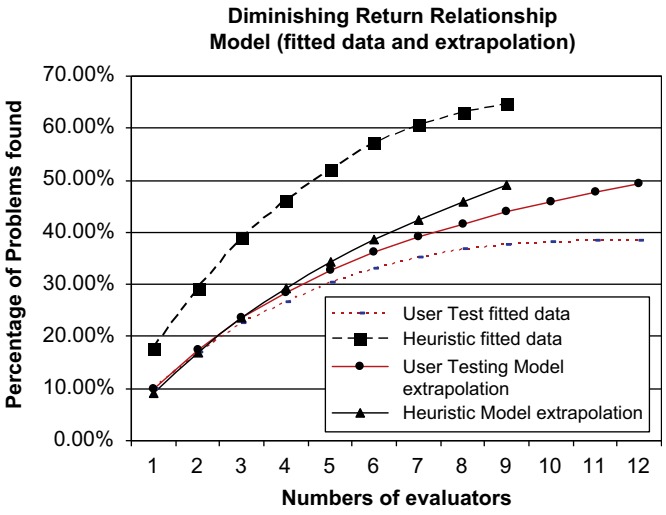

Fig. 5. Plot of the developed model extrapolation data vs. fitted data.

4. Discussion

Every user interface has its fair share of usability problems. Unfortunately, the exact number and type of problems are not known. Web site evaluation methods such as user testing and heuristic analysis seek to solve this dilemma by detecting and predicting these usability problems. With every usability inspection, a little bit more is learned about the user interface. Some of the usability problems can easily be detected, while some may be harder to detect.

The primary intent of this study was to compare the efficiency and effectiveness of user testing and heuristic analysis in detecting and predicting usability problems. From this study it can be concluded that:

- Both user testing and heuristic methods compliment each other. Neither one method can be replaced by the other.
- The two methods address different problems and different levels of severity. Due to the different techniques used in the two methods, the content and the nature of the problems found by these two methods vary to some extent. Heuristic evaluation tends to cover more high-level structural problems and likely to address some of the root causes of these problems, because evaluators have the flexibility to assess every aspect of the feature, while user testing is solely dependent on the predefined scenarios. If the scenario did not include a possibility of failure, then this problem remains salient from user testing. But on the other hand, user testing can reveal some detail levels of a certain design feature because user is required to get to the task level to complete it. For example, in heuristic evaluation, evaluators identified problems such as "Site tools provided at the bottom are too many", "Page length not printer friendly," "Information tool should be a narrow search when particular item is viewed", but failed to identify problems such as "Unable to locate submit button for shipping information", "Search result list items not appropriate based not keyword (vitamin B)" etc., but rather easily reported from user testing. In our study, heuristic evaluation covered a wider range of problems than the user testing. This can be seen from the distribution of the seven categories of the problems found from these two methods.
- Heuristic analysis identifies more number of problems than user testing.
- Given that web development goes through different stages, from initial concept to final construction, these methods may have better, individual validity at different stages.
- This study found that 5 subjects or evaluators could find only 35% of the usability problems in a user interface. This is not consistent with results reported by Nielsen and Mack (1994). There, Nielsen and Mack (1994) reported that 5 users are adequate to address as many as 85% of the usability problems on a web site. One possible explanation may be attributed to the fact that Nielsen and Mack (1994) had not assumed the total number of problems in a web site to be the sum of problems addressed by both the user testing and heuristic. Another possible reason could be the definition of experts in heuristic evaluation. Some evaluators used in this study might not be qualified as a true expert so their ability of finding more comprehensive usability problems could be limited.
- The result from this study is more consistent with Spool and Schroeder (2001), who estimated that 5 users could only find 35% of the problems.
- For heuristic analysis, the numbers of evaluators required are dependent on two important factors, which include the heuristic used, and the level of experience of the heuristic evaluators. For example, it can be suggested that given a more detailed set of heuristics or a more seasoned evaluator, it may take fewer evaluators to address the same amount of usability problems. Although this study did not address this issue specifically, but based on data and observation from this study, it is believed that 2–3 highly experienced evaluators, or 3–5 intermediately experienced evaluators will be adequate to evaluate an interface.

In summary, our recommendations are that both user testing and heuristic analysis are needed in a usability study. This is consistent with other studies that have consistently shown that different methods have different strengths; the best evaluation of a user interface comes from applying multiple evaluation techniques (Jeffries et al., 1991; Desurvire et al., 1992). However, in order to reap the optimal benefits, both user testing and heuristic analysis should be used in different stages of the user interface design process. We believe that heuristic analysis should be implemented at early stages of the development process, while user testing should be conducted at a later stage of the development process.

It is believed that the difference in nature of these two techniques would make them appropriate for different testing purposes. In our study, heuristic analysis finds more problems than user testing because it provides more freedom exploring the interface, while user testing needs a well-developed test bed and in a more controlled environment (Preece et al., 2002). Typically at the earlier design stage, the interface is often not fully developed. Heuristic analysis would be able to project potential usability problems, a quality that user testing lacks. Feedbacks from heuristic analysis can be used to create a design standard for the rest of the web site. After design improvements are made, following the initial heuristic analysis, a thorough user testing is required as user testing and heuristic analysis finds very different and specific types of problems. User testing would be able to assess the usability issues most pertinent to users much more directly, without bothering with the basic problems. Feedback from user testing can be used to fine-tune the web site, which is typically done at the later stage of the design process. User testing may also detect potential new usability problems that were the direct result of the design improvement.

In conclusion, our study showed that both user testing and heuristic analysis are needed in a usability study. In order to reap the optimal benefits, it is believed that both user testing and heuristic analysis should be used in different stages of the user

interface design process. Since web pages have content and usage that are specific to that domain, one can only make high-level generalization from this study. More study, specific to the web page of interest would be needed to make specific recommendations.

References

Desurvire, H.W., Kondziela, J.M., Atwood, M.E., 1992. What is gained and lost when using evaluation methods other than empirical testing 89–102. In: Monk, A., Diaper, D., Harrison, M.D. (Eds.), People and Computers VII. Cambridge University Press, Cambridge, pp. 89–102, A Shorter Version of this Paper is Available in the Digest of Shot Talks Presented at CHI'92, Monterey, CA, May 7, pp. 125–126.

Doubleday, A., Ryan, M. Springett, M., Sutcliffe, A., 1997. A comparison of usability techniques for evaluating design. In: Proceedings of DIS 97. ACM, New York, NY, pp. 101–110.

Gramopadhye, A.K., Drury, C.G., 2000. Human factors in aviation maintenance: how we got to where we are. International Journal of Industrial Ergonomics 26 (2), 125–131.

Jeffries, R.J., Miller, J.R., Wharton, C., Uyeda, K.M., 1991. User interface evaluation in the real world: a comparison of four techniques. In: Proceedings ACM CHI'91 Conference, New Orleans, LA, April 29–May 2, 1991, pp. 119–124.

Latorella, K.A, Prabhu, P.V, 2000. A review of human error in aviation maintenance and inspection. International Journal of Industrial Ergonomics 26 (2), 133–161.

Liljegren, E., 2006. Usability in a medical technology context assessment of methods for usability evaluation of medical equipment. International Journal of Industrial Ergonomics 36 (4), 345–352.

Liljegren, E, Osvalder, A.L, 2004. Cognitive engineering methods as usability evaluation tools for medical equipment. International Journal of Industrial Ergonomics 34 (1), 49–62.

Lindgaard, G, 2006. Notions of thoroughness, efficiency, and validity: are they valid in HCI practice? International Journal of Industrial Ergonomics 36 (12), 1069–1074.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., Kirakowski, J., 1998. Comparative evaluation of usability tests. In: The Proceeding of 1998 Usability Professionals' Association Conference, Capitalizing on Usability, Washington, DC, USA, June 22–26, 1998.

Nielsen, J. Molich, R., 1990. Heuristic evaluation of user interfaces. In: Proceedings of CHI 90. ACM, New York, NY, pp. 249–256.

Nielsen, J., Landauer, T.K. 1993. A mathematical model of the finding of usability problems. In: Proceedings of ACM INTERCHI'93 Conference, Amsterdam, The Netherlands, April 24–29, 1993, pp 206–213.

Nielsen, J., Mack, R.L., 1994. Usability Inspection Methods. Wiley, New York.

Preece, J., Rogers, Y., Sharp, H., 2002. Interaction Design: Beyond Human–Computer Interaction. Wiley, New York, NY.

Rexfelt, O., Rosenblad, E., 2006. The progress of user requirements through a software development project. International Journal of Industrial Ergonomics 36 (1), 73–81.

Spool, J. Schroeder, W., 2001. Testing Websites: Five Users is Nowhere Near Enough. In: Proceedings CHI 2001, Extended Abstracts, ACM pp. 285–286.

Tan, W., 2003. A comparison of user testing and heuristic analyses of web sites. An unpublished MS Thesis, Department of Industrial and Management Systems Engineering, University of Nebraska, May 2003.

Woolrych, A. Cockton, G., 2002. Testing a conjecture based on the DR-AR model of usability inspection method effectiveness. In: Sharp, H., et al. (Eds.), Proceeding of HCI 2002 Conference, 2. British Computer Society, London, 2002.